Decomposable Graphical Gaussian Models

Steffen Lauritzen, University of Oxford

CIMPA Summerschool, Hammamet 2011, Tunisia

September 12, 2011
This simple algorithm has complexity $O(|V| + |E|)$:

1. Choose $v_0 \in V$ arbitrary and let $v_0 = 1$;
2. When vertices $\{1, 2, \ldots, j\}$ have been identified, choose $v = j + 1$ among $V \setminus \{1, 2, \ldots, j\}$ with highest cardinality of its numbered neighbours;
3. *If* $\text{bd}(j + 1) \cap \{1, 2, \ldots, j\}$ *is not complete*, $\mathcal{G}$ is not chordal;
4. Repeat from 2;
5. *If the algorithm continues until no vertices are left, the graph is chordal and the numbering is perfect.*
Finding the cliques of a chordal graph

From an MCS numbering $V = \{1, \ldots, |V|\}$, let

$$B_\lambda = \text{bd}(\lambda) \cap \{1, \ldots, \lambda - 1\}$$

and $\pi_\lambda = |B_\lambda|$. Call $\lambda$ a *ladder vertex* if $\lambda = |V|$ or if $\pi_{\lambda+1} < \pi_\lambda + 1$. Let $\Lambda$ be the set of ladder vertices.

The cliques are $C_\lambda = \{\lambda\} \cup B_\lambda$, $\lambda \in \Lambda$. 

[Diagram of a chordal graph with vertices labeled 1 to 7 and edges connecting them.]

$\pi_\lambda$: 0, 1, 2, 2, 2, 1, 1.
Let $\mathcal{A}$ be a collection of finite subsets of a set $V$. A *junction tree* $T$ of sets in $\mathcal{A}$ is an undirected tree with $\mathcal{A}$ as a vertex set, satisfying the *junction tree property*:

*If $A, B \in \mathcal{A}$ and $C$ is on the unique path in $T$ between $A$ and $B$ it holds that $A \cap B \subset C$."

If the sets in an arbitrary $\mathcal{A}$ are pairwise incomparable, *they can be arranged in a junction tree if and only if $\mathcal{A} = \mathcal{C}$ where $\mathcal{C}$ are the cliques of a chordal graph.*
The following are equivalent for any undirected graph $G$.

(i) \( G \) is chordal;
(ii) \( G \) is decomposable;
(iii) All prime components of \( G \) are cliques;
(iv) \( G \) admits a perfect numbering;
(v) Every minimal \((\alpha, \beta)\)-separator are complete.
(vi) The cliques of \( G \) can be arranged in a junction tree.
The junction tree can be constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$, where $C_\lambda$ are the cliques: Since the MCS-numbering is perfect, $C_\lambda, \lambda > \lambda_{\min}$ all satisfy

$$C_\lambda \cap (\bigcup_{\lambda' < \lambda} C_{\lambda'}) = C_\lambda \cap C_{\lambda^*} = S_\lambda$$

for some $\lambda^* < \lambda$.

A junction tree is now easily constructed by attaching $C_\lambda$ to any $C_{\lambda^*}$ satisfying the above. Although $\lambda^*$ may not be uniquely determined, $S_\lambda$ is.

Indeed, the sets $S_\lambda$ are the minimal complete separators and the numbers $\nu(S)$ are $\nu(S) = |\{\lambda \in \Lambda : S_\lambda = S\}|$. 
A chordal graph
Junction tree

Cliques of graph arranged into a tree with $C_1 \cap C_2 \subseteq D$ for all cliques $D$ on path between $C_1$ and $C_2$. 
In general, the *prime components* of any undirected graph can be arranged in a junction tree in a similar way.

Then *every pair of neighbours (C, D) in the junction tree represents a decomposition of G* into $G_{\tilde{C}}$ and $G_{\tilde{D}}$, where $\tilde{C}$ is the set of vertices in prime components connected to $C$ but separated from $D$ in the junction tree, and similarly with $\tilde{D}$.

The corresponding algorithm is based on a slightly more sophisticated algorithm known as *Lexicographic Search* (LEX) which runs in $O(|V|^2)$ time.
If the graph $\mathcal{G}$ is chordal, we say that the graphical model is *decomposable*.

In this case, the IPS-algorithm converges in a finite number of steps.

We also have the familiar *factorization of densities*

\[
f(x \mid \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \Sigma_S)^{\nu(S)}} \tag{1}
\]

where $\nu(S)$ is the number of times $S$ appear as intersection between neighbouring cliques of a junction tree for $\mathcal{C}$.
Relations for trace and determinant

Using the factorization (1) we can for example match the expressions for the trace and determinant of $\Sigma$

$$\text{tr}(KW) = \sum_{C \in C} \text{tr}(K_C W_C) - \sum_{S \in S} \nu(S) \text{tr}(K_S W_S)$$

and further

$$\det \Sigma = \{\det(K)\}^{-1} = \frac{\prod_{C \in C} \det\{\Sigma_C\}}{\prod_{S \in S} \{\det(\Sigma_S)\}^{\nu(S)}}$$

These are some of many relations that can be derived using the decomposition property of chordal graphs.
The same factorization clearly holds for the maximum likelihood estimates:

\[
f(x | \hat{\Sigma}) = \frac{\prod_{C \in C} f(x_C | \hat{\Sigma}_C)}{\prod_{S \in S} f(x_S | \hat{\Sigma}_S)^{\nu(S)}}
\]  

(2)

Moreover, it follows from the general likelihood equations that

\[
\hat{\Sigma}_A = \mathcal{W}_A / n \text{ whenever } A \text{ is complete.}
\]

Exploiting this, we can obtain an explicit formula for the maximum likelihood estimate in the case of a chordal graph.
For a \(|d| \times |e|\) matrix \(A = \{a_{\gamma \mu}\}_{\gamma \in d, \mu \in e}\) we let \([A]^V\) denote the matrix obtained from \(A\) by filling up with zero entries to obtain full dimension \(|V| \times |V|\), i.e.

\[
([A]^V)_{\gamma \mu} = \begin{cases} 
  a_{\gamma \mu} & \text{if } \gamma \in d, \mu \in e \\
  0 & \text{otherwise.}
\end{cases}
\]

The maximum likelihood estimates exists if and only if \(n \geq C\) for all \(C \in \mathcal{C}\). Then the following simple formula holds for the maximum likelihood estimate of \(K\):

\[
\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} \left[ (w_C)^{-1} \right]^V - \sum_{S \in S} \nu(S) \left[ (w_S)^{-1} \right]^V \right\}.
\]

Steffen Lauritzen, University of Oxford

Decomposable Graphical Gaussian Models
This graph is chordal with cliques \( \{1, 2, 3\} \), \( \{3, 4, 5\} \) with separator \( S = \{3\} \) having \( \nu(\{3\}) = 1 \).
Since one degree of freedom is lost by subtracting the average, we get in this example

$$\hat{K} = 87 \begin{pmatrix} w_{[123]}^{11} & w_{[123]}^{12} & w_{[123]}^{13} & 0 & 0 \\ w_{[123]}^{21} & w_{[123]}^{22} & w_{[123]}^{23} & 0 & 0 \\ w_{[123]}^{31} & w_{[123]}^{32} & w_{[123]}^{33} + w_{[345]}^{33} - 1/w_{33} & w_{[345]}^{34} & w_{[345]}^{35} \\ 0 & 0 & w_{[345]}^{43} & w_{[345]}^{44} & w_{[345]}^{45} \\ 0 & 0 & w_{[345]}^{53} & w_{[345]}^{54} & w_{[345]}^{55} \end{pmatrix}$$

where $w_{[123]}^{ij}$ is the $ij$th element of the inverse of

$$W_{[123]} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix}$$

and so on.
The formula

\[ \hat{\Sigma} = n^{-1} \left\{ \sum_{C \in \mathcal{C}} \left( (W_C)^{-1} \right)^V - \sum_{S \in \mathcal{S}} \nu(S) \left( (W_S)^{-1} \right)^V \right\}^{-1} \]

specifies \( \hat{\Sigma} \) as a random matrix.

The distribution of this random Wishart-type matrix is partly reflecting Markov properties of the graph \( \mathcal{G} \).

This is also true for the distribution of \( \hat{\Sigma} \) for a non-chordal graph \( \mathcal{G} \) but not to the same degree.

Before we delve further into this, we shall need some more terminology.
Laws and distributions

Generally we think of

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$

and sometimes identify $\Theta$ with $\mathcal{P}$ which is justified when the parametrization

$$\theta \rightarrow P_\theta$$

is one-to-one and onto.

In a Gaussian graphical model $\theta = K \in S^+(\mathcal{G})$ is uniquely identifying any regular Gaussian distribution satisfying the Markov properties w.r.t. $\mathcal{G}$. 
In any case, any probability measure on $\mathcal{P}$ (or on $\Theta$) represents a random element of $\mathcal{P}$, i.e. a random distribution. The *sampling distribution of the MLE $\hat{\Sigma}$* is an example of such a measure.

To keep heads straight we refer to a probability measure on $\mathcal{P}$ as a *law*, whereas a *distribution* is a probability measure on $\mathcal{X}$.

Thus we shall e.g. speak of the *Wishart law* as we think of it specifying a distribution of $f(\cdot | \Sigma)$. 
We identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. $\theta_A$ for $A \subseteq V$ denotes the distribution of $X_A$ under $P_\theta$ and $\theta_A|B$ the family of conditional distributions of $X_A$ given $X_B$, etc.

For a law $\mathcal{L}$ on $\Theta$ we write

$$A \perp \!\!\! \perp \mathcal{L} B \mid S \iff \theta_{A \cup S} \perp \!\!\! \perp \mathcal{L} \theta_{B \cup S} \mid \theta_S.$$ 

A law $\mathcal{L}$ on $\Theta$ is hyper Markov w.r.t. $G$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $G$;

(ii) $A \perp \!\!\! \perp \mathcal{L} B \mid S$ whenever $S$ is complete and $A \perp G B \mid S$.

Note the conditional independence is only required to hold for graph decompositions.
If $\theta$ follows a hyper Markov law for this graph, it holds for example that

$$\theta_{1235} \independent \independent \theta_{24567} \mid \theta_{25}.$$ 

_This is indeed true for $\hat{\Sigma}$ in the graphical model with this graph_, i.e.

$$\hat{\Sigma}_{1235} \independent \independent \hat{\Sigma}_{24567} \mid \hat{\Sigma}_{25}.$$
Consequences of the hyper Markov property

Clearly, if $A \perp \perp B \mid S$, we have for example also (using property (C2) of conditional independence)

$$\theta_A \perp \perp \theta_B \mid \theta_S$$

since $\theta_A$ and $\theta_B$ are functions of $\theta_{A \cup S}$ and $\theta_{B \cup S}$ respectively. But the converse is false! $\theta_A \perp \perp \theta_B \mid \theta_S$ does not imply $\theta_{A \cup S} \perp \perp \theta_{B \cup S} \mid \theta_S$, since $\theta_{A \cup S}$ is not a function of $(\theta_A, \theta_S)$. In contrast, $X_{A \cup B}$ is indeed a (one-to-one) function of $(X_A, X_B)$.

However it generally holds that

$$A \perp \perp B \mid S \iff \theta_A \mid S \perp \perp \theta_B \mid S \mid \theta_S.$$
Chordal graphs

If $G$ is chordal and $\theta$ is hyper Markov on $G$, it holds that

$$A \perp_G B \mid S \Rightarrow A \perp_{\mathcal{L}} B \mid S$$

i.e. it is not necessary to specify that $S$ is a complete separator to obtain the relevant conditional independence.

This follows essentially because for a chordal graph it holds that

$$A \perp_G B \mid S \Rightarrow \exists S^* \subseteq S : A \perp_G B \mid S^* \text{ with } S^* \text{ complete.}$$

If $G$ is not chordal, we can form $\overline{G}$ by completing all prime components of $G$.

If $\theta$ is hyper Markov on $G$ it is also hyper Markov on $\overline{G}$, and thus

$$A \perp_{\overline{G}} B \mid S \Rightarrow A \perp_{\mathcal{L}} B \mid S.$$

But the similar result is \textit{false} for an arbitrary chordal cover of $G$. 
Directed hyper Markov property

We have similar notions and results in the directed case. Say $\mathcal{L} = \mathcal{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathcal{D}$ if $\theta$ is directed Markov on $\mathcal{D}$ for all $\theta \in \Theta$ and

$$
\theta_{v \cup \text{pa}(v)} \perp \perp \mathcal{L} \theta_{\text{nd}(v)} \mid \theta_{\text{pa}(v)},
$$

or equivalently $\theta_{v \mid \text{pa}(v)} \perp \perp \mathcal{L} \theta_{\text{nd}(v)} \mid \theta_{\text{pa}(v)}$, or equivalently for a well-ordering

$$
\theta_{v \cup \text{pa}(v)} \perp \perp \mathcal{L} \theta_{\text{pr}(v)} \mid \theta_{\text{pa}(v)}.
$$

In general there is no similar statement corresponding to the global property and $d$-separation. However, if $\mathcal{D}$ is perfect, $\mathcal{L}$ is directed hyper Markov w.r.t. $\mathcal{D}$ if and only if $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$. 
The distributions of maximum likelihood estimators are important examples of hyper Markov laws. But for \textit{chordal graphs} there is a canonical construction of such laws.

Let $\mathcal{C}$ be the cliques of a chordal graph $\mathcal{G}$ and let $\mathcal{L}_C, C \in \mathcal{C}$ be a family of laws over $\Theta_C \subseteq \mathbb{P}(\mathcal{X}_C)$.

The family of laws are \textit{hyperconsistent} if for any $C$ and $D$ with $C \cap D = S \neq \emptyset$, $\mathcal{L}_C$ and $\mathcal{L}_D$ induce the same law for $\theta_S$.

\textit{If $\mathcal{L}_C, C \in \mathcal{C}$ are hyperconsistent, there is a unique hyper Markov law $\mathcal{L}$ over $\mathcal{G}$ with $\mathcal{L}(\theta_C) = \mathcal{L}_C$, $C \in \mathcal{C}$.}
In some cases it is of interest to consider a stronger version of the hyper Markov property.

A hyper Markov law is *strongly hyper Markov* if $\theta_A | S \perp \perp \theta_S$ for all complete separators $S$.

A directed hyper Markov law is *strongly directed hyper Markov* if $\theta_v | pa(v) \perp \perp \theta_{pa(v)}$ for all $v \in V$. 
Parameter $\theta \in \Theta$, data $X = x$, likelihood

$$L(\theta \mid x) \propto p(x \mid \theta) = \frac{dP_\theta(x)}{d\mu(x)}.$$ 

Express knowledge about $\theta$ through a prior law $\pi$ on $\theta$. Use also $\pi$ to denote density of the prior law w.r.t. some measure $\nu$ on $\Theta$. Inference about $\theta$ from $x$ is then represented through posterior law $\pi^*(\theta) = p(\theta \mid x)$. Then, from Bayes’ formula

$$\pi^*(\theta) = p(x \mid \theta)\pi(\theta)/p(x) \propto L(\theta \mid x)\pi(\theta)$$

so the likelihood function is equal to the density of the posterior w.r.t. the prior modulo a constant.
A family $\mathcal{P}$ of laws on $\Theta$ is said to be \textit{stable under sampling} from $x$ if

$$\pi \in \mathcal{P} \Rightarrow \pi^* \in \mathcal{P}.$$ 

If the family of priors is parametrised:

$$\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$$

we sometimes say that $\alpha$ is a \textit{hyperparameter}. Then, Bayesian inference can be made by just updating hyperparameters.
If $\mathcal{L}$ is a prior law over $\Theta$ and $X = x$ is an observation from $\theta$, $\mathcal{L}^* = \mathcal{L}(\theta | X = x)$ denotes the posterior law over $\Theta$.

If $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

If $\mathcal{L}$ is strongly hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

In the latter case, the update of $\mathcal{L}$ is local to prime components, i.e.

$$\mathcal{L}^*(\theta_Q) = \mathcal{L}_Q^*(\theta_Q) = \mathcal{L}_Q(\theta_Q | X_Q = x_Q)$$

and the marginal distribution $p$ of $X$ is globally Markov w.r.t. $\overline{\mathcal{G}}$, where

$$p(x) = \int_{\Theta} P(X = x | \theta) \mathcal{L}(d\theta).$$
For a \( k \)-dimensional exponential family

\[
p(x \mid \theta) = b(x) e^{\theta^\top t(x) - \psi(\theta)}
\]

the \textit{standard conjugate family} is given as

\[
\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa \psi(\theta)}
\]

for \((a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+\), where \(\mathcal{A}\) is determined so that the normalisation constant is finite.

\textit{The conjugate family is stable under sampling} and posterior updating from \((x_1, \ldots, x_n)\) with \(t = \sum_i t(x_i)\) is then made as \((a^*, \kappa^*) = (a + t, \kappa + n)\).
Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

\[ \pi(K | \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\text{tr}(K\Phi)}, \quad K \in S^+(G). \]

These laws are termed *hyper inverse Wishart laws* as \( \Sigma \) follows an inverse Wishart law for complete graphs.

*For chordal graphs, each marginal law \( \mathcal{L}_C \) of \( \Sigma_C \) is inverse Wishart.*